GDP vs Suicide Rates over the Past 30 Years

Group3
-ChungHyun Lee
-Evgeniy Ko
-Tek Acharya

Abstract

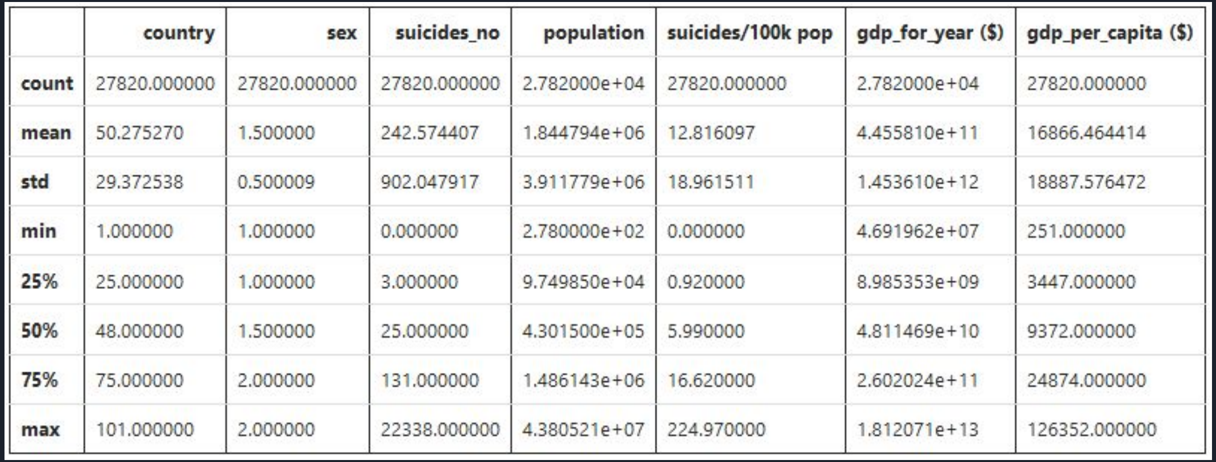
- Suicide represents about ~1.5% of deaths in the U.S and in the world
- Is there a correlation between GDP and suicide rates
- K-Means will help us identify subgroups in our data
- PCA will help us reduce the dimensionality of our large data set
- Linear regression will provide us a regression function

Data set

- Suicide rates in 101 different countries, over 30 years

- 6 categorical variables: country, year, sex, age, country-year, and generation

- 6 numerical variables: suicide number, population, suicides/100k population, HDI for year, GDP for that year, and GDP per capita
  HDI has a lot of missing values



	country	sex	suicides_no	population	suicides/100k pop	gdp_for_year ($)	gdp_per_capita ($)
count	27820.000000	27820.000000	27820.000000	2.782000e+04	27820.000000	2.782000e+04	27820.000000
mean	50.275270	1.500000	242.574407	1.844794e+06	12.816097	4.455810e+11	16866.464414
std	29.372538	0.500009	902.047917	3.911779e+06	18.961511	1.453610e+12	18887.576472
min	1.000000	1.000000	0.000000	2.780000e+02	0.000000	4.691962e+07	251.000000
25%	25.000000	1.000000	3.000000	9.749850e+04	0.920000	8.985353e+09	3447.000000
50%	48.000000	1.500000	25.000000	4.301500e+05	5.990000	4.811469e+10	9372.000000
75%	75.000000	2.000000	131.000000	1.486143e+06	16.620000	2.602024e+11	24874.000000
max	101.000000	2.000000	22338.000000	4.380521e+07	224.970000	1.812071e+13	126352.000000

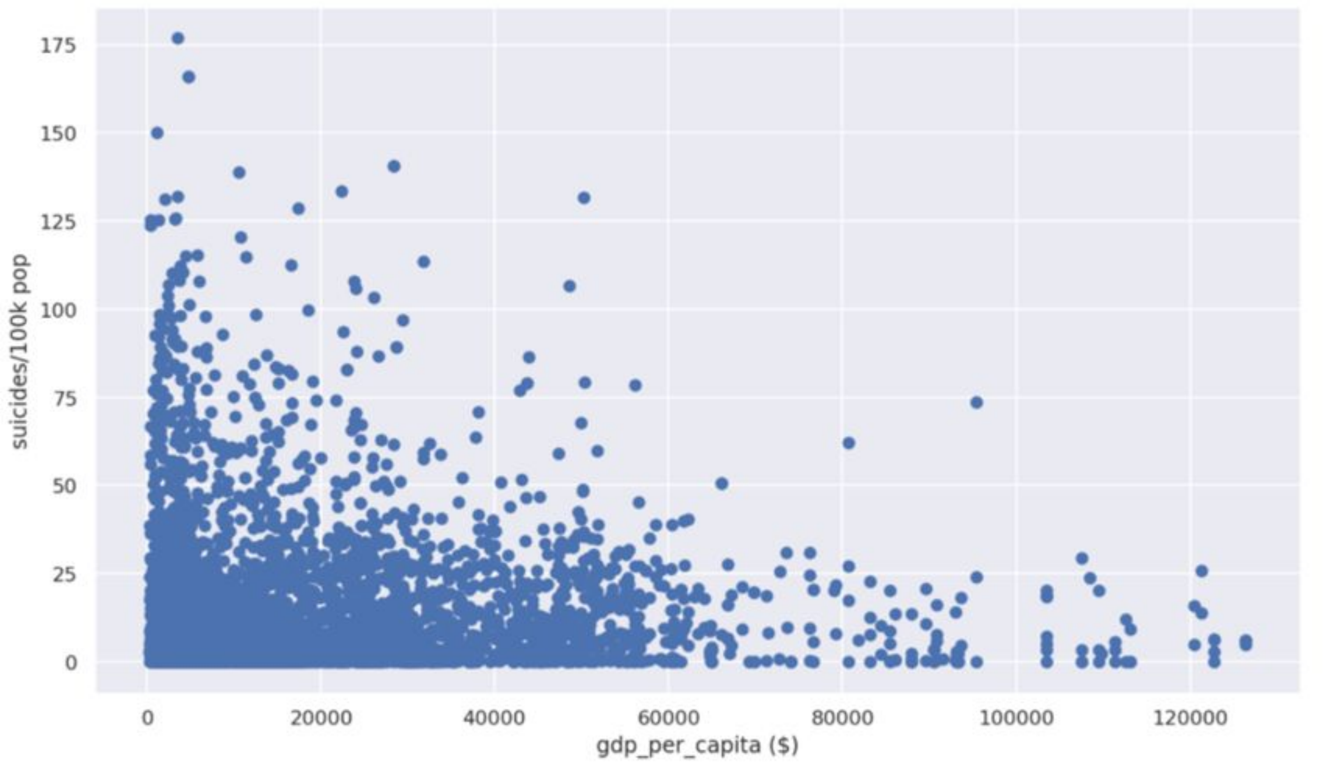
Predictions

Chunghyun Lee: A strong positive correlation between GDP and suicide rates

Tek Acharya: A strong positive relationship between GDP and suicide rates as country's economy has direct impact on people's lives

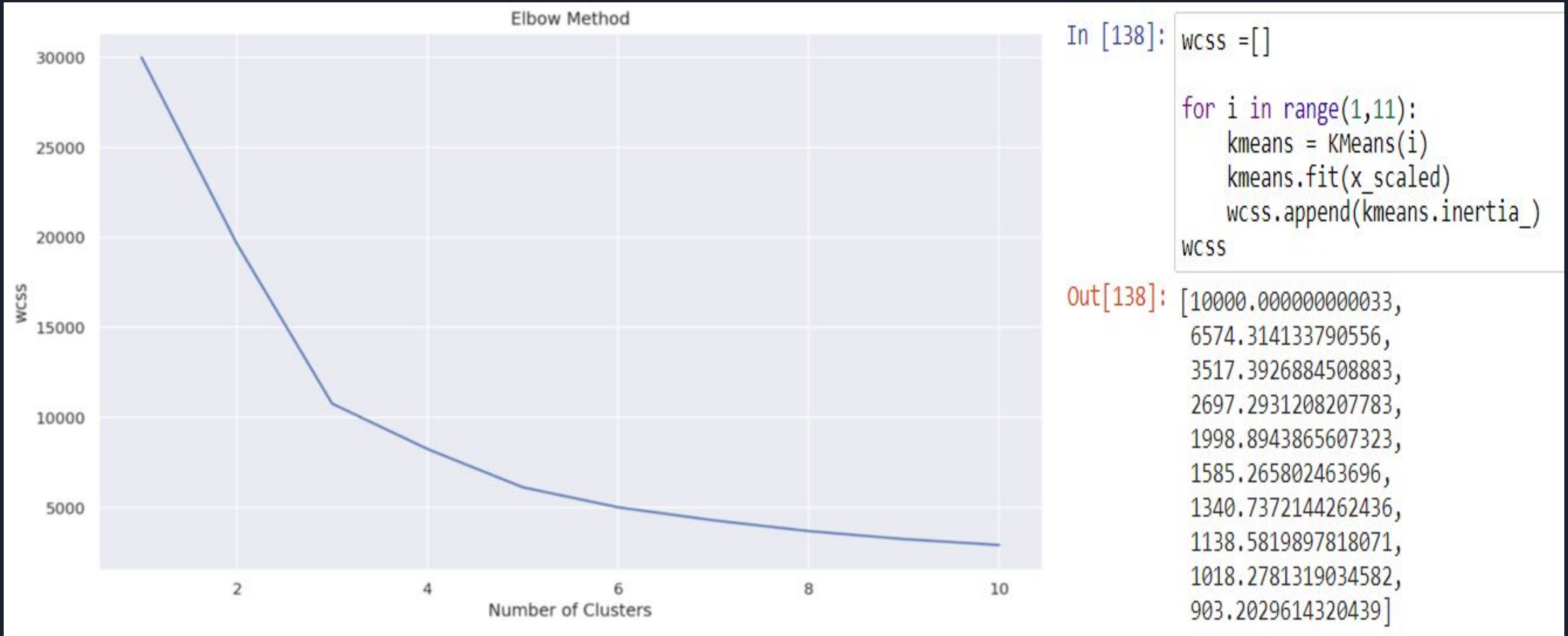
Evgeniy Ko: A weak correlation between GDP and suicide rates, because there are other factors that effect suicide rates

K-Means



suicides/100k pop
175
150
125
100
75
50
25
0
0
20000
40000
60000
80000
100000
120000
gdp_per_capita ($)

K-Means



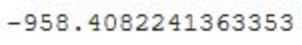
```
In [138]: wcss =[]

          for i in range(1,11):
              kmeans = KMeans(i)
              kmeans.fit(x_scaled)
              wcss.append(kmeans.inertia_)
          wcss
```

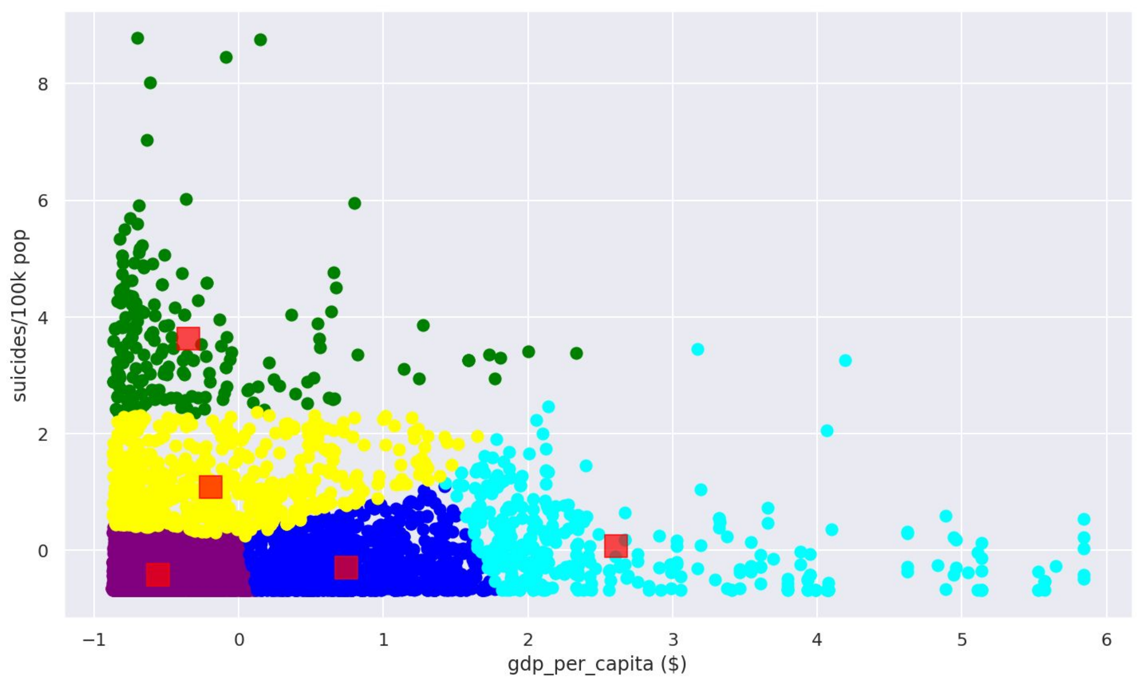
```
Out[138]: [10000.000000000033,
           6574.314133790556,
           3517.3926884508883,
           2697.2931208207783,
           1998.8943865607323,
           1585.265802463696,
           1340.7372144262436,
           1138.5819897818071,
           1018.2781319034582,
           903.2029614320439]
```

K-Means

K-Means Algorithm Score:

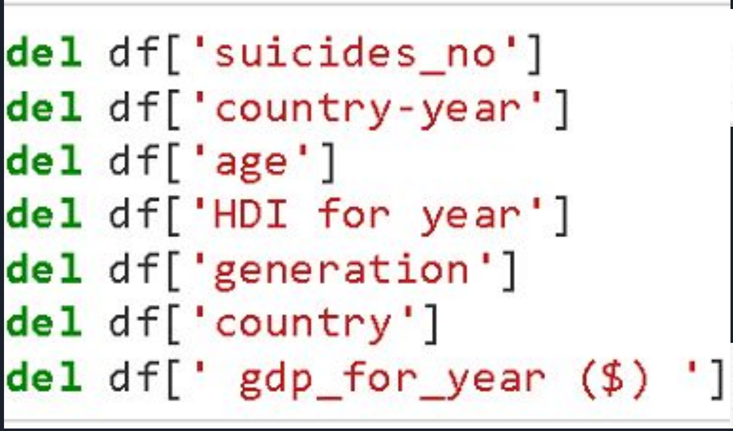


```
-958.4082241363353
```

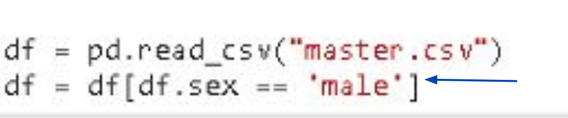


PCA

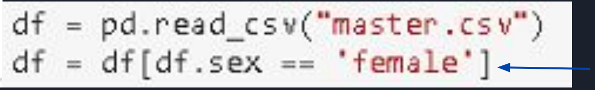
- We decided to go with “Year”, “Population”, “suicide/100k pop” and “gdp_per_capita” as we believed that rest of the features do not contribute to suicide rate.



```
del df['suicides_no']
del df['country-year']
del df['age']
del df['HDI for year']
del df['generation']
del df['country']
del df[' gdp_for_year ($) ']
```



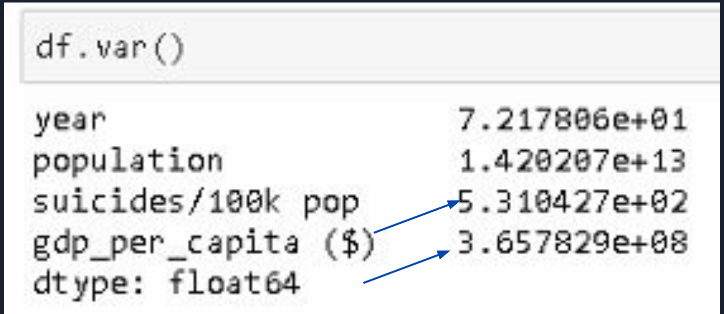
```
df = pd.read_csv("master.csv")
df = df[df.sex == 'male']
```



```
df = pd.read_csv("master.csv")
df = df[df.sex == 'female']
```

PCA

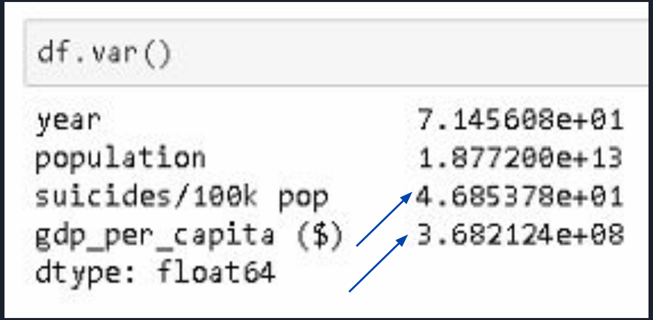
Male



```
df.var()
```

```
year                    7.217806e+01
population              1.420207e+13
suicides/100k pop       5.310427e+02
gdp_per_capita ($)      3.657829e+08
dtype: float64
```

Female

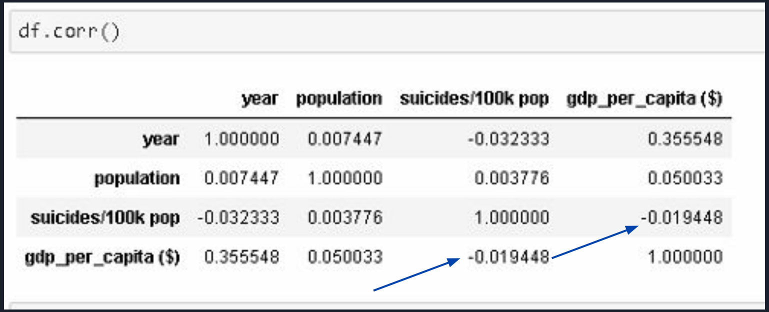


```
df.var()
```

```
year                    7.145608e+01
population              1.877200e+13
suicides/100k pop       4.685378e+01
gdp_per_capita ($)      3.682124e+08
dtype: float64
```

PCA

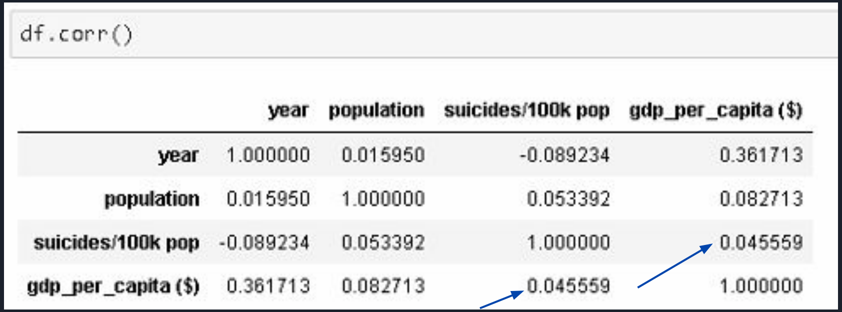
Male



```
df.corr()
```

	year	population	suicides/100k pop	gdp_per_capita ($)
year	1.000000	0.007447	-0.032333	0.355548
population	0.007447	1.000000	0.003776	0.050033
suicides/100k pop	-0.032333	0.003776	1.000000	-0.019448
gdp_per_capita ($)	0.355548	0.050033	-0.019448	1.000000

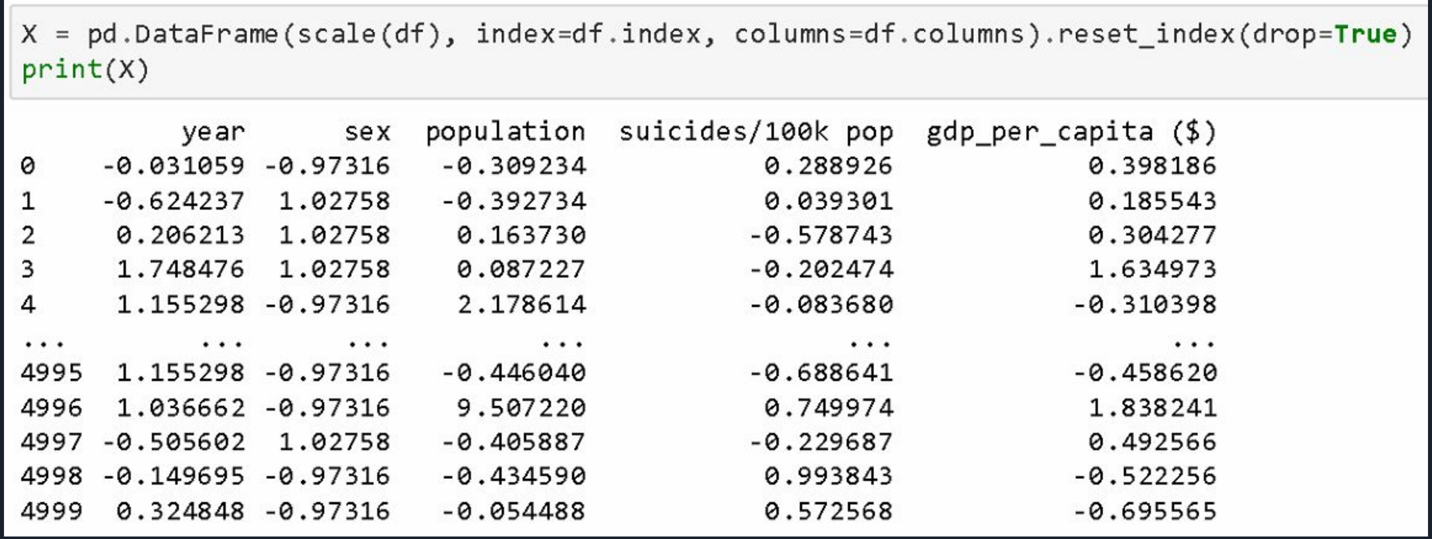
Female



```
df.corr()
```

	year	population	suicides/100k pop	gdp_per_capita ($)
year	1.000000	0.015950	-0.089234	0.361713
population	0.015950	1.000000	0.053392	0.082713
suicides/100k pop	-0.089234	0.053392	1.000000	0.045559
gdp_per_capita ($)	0.361713	0.082713	0.045559	1.000000

PCA



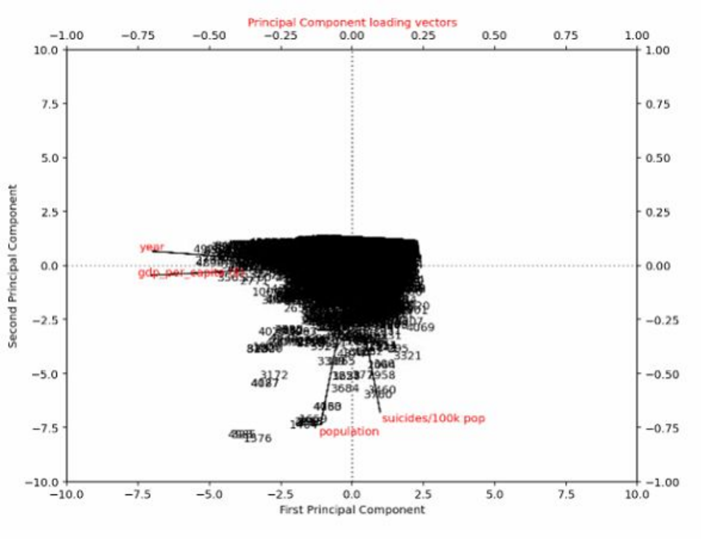
```
X = pd.DataFrame(scale(df), index=df.index, columns=df.columns).reset_index(drop=True)
print(X)
```

```
          year      sex  population  suicides/100k pop  gdp_per_capita ($)
0    -0.031059 -0.97316   -0.309234           0.288926            0.398186
1    -0.624237  1.02758   -0.392734           0.039301            0.185543
2     0.206213  1.02758    0.163730          -0.578743            0.304277
3     1.748476  1.02758    0.087227          -0.202474            1.634973
4     1.155298 -0.97316    2.178614          -0.083680           -0.310398
...        ...      ...         ...                ...                 ...
4995  1.155298 -0.97316   -0.446040          -0.688641           -0.458620
4996  1.036662 -0.97316    9.507220           0.749974            1.838241
4997 -0.505602  1.02758   -0.405887          -0.229687            0.492566
4998 -0.149695 -0.97316   -0.434590           0.993843           -0.522256
4999  0.324848 -0.97316   -0.054488           0.572568           -0.695565
```

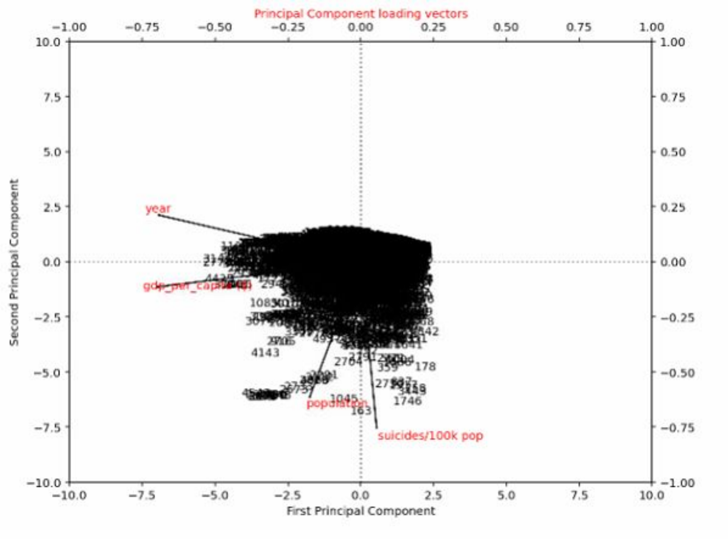
PCA

- Looks to us that the PCA (plot) for both male and female to be very similar where PC1 carries little more information to that of PC2

Male



Female



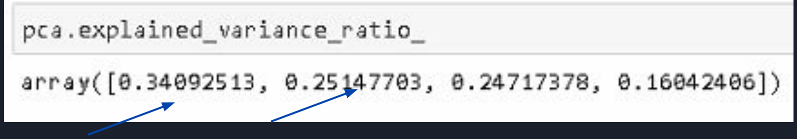
PCA

Male

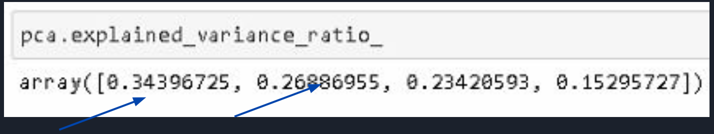
- The PC1 explained: 34.09%
- The PC2 explained: 25.14%

Female

- PC1 explained: 34.40%
- PC2 explained: 26.90%



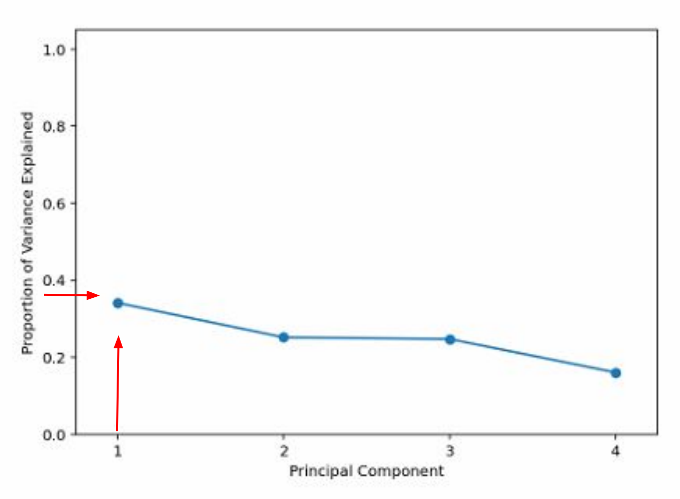
```
pca.explained_variance_ratio_
array([0.34092513, 0.25147703, 0.24717378, 0.16042406])
```



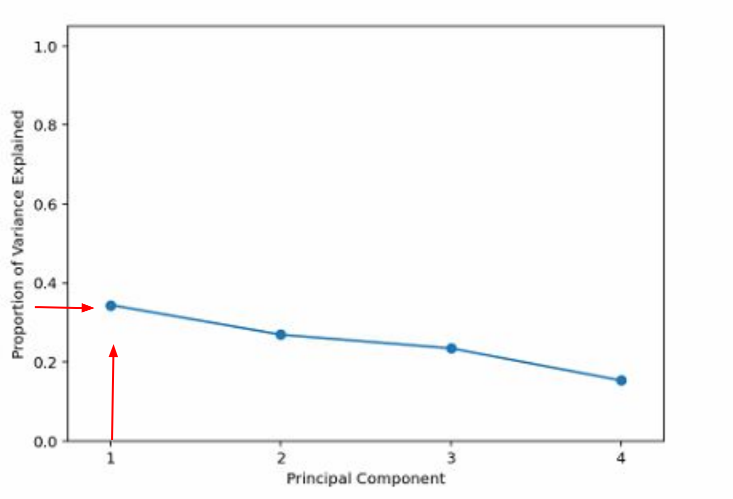
```
pca.explained_variance_ratio_
array([0.34396725, 0.26886955, 0.23420593, 0.15295727])
```

PCA

Male

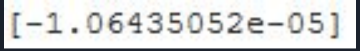


Female



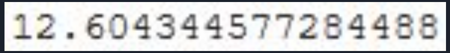
Linear Regression

- For the first model, we performed the algorithm on a sample, n=5000, of the original data set.
- Model’s Coefficient

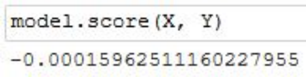
  ```
  [-1.06435052e-05]
  ```

- Model’s Intercept

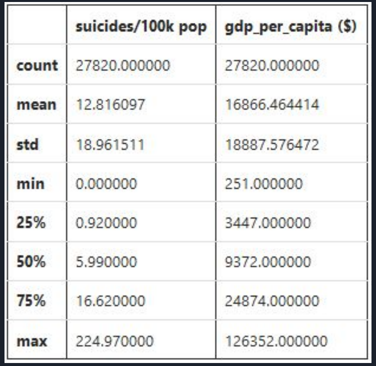
  ```
  12.604344577284488
  ```

- R-Squared

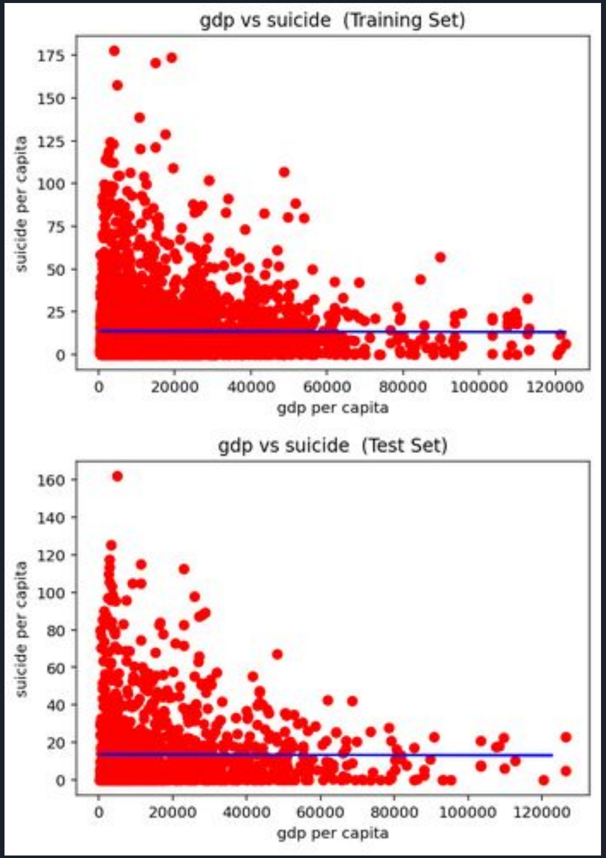
  

  ```
  model.score(X, Y)
  ```

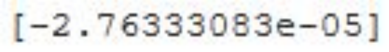
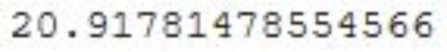
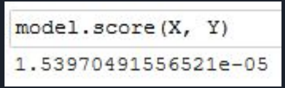
  ```
  -0.00015962511160227955
  ```

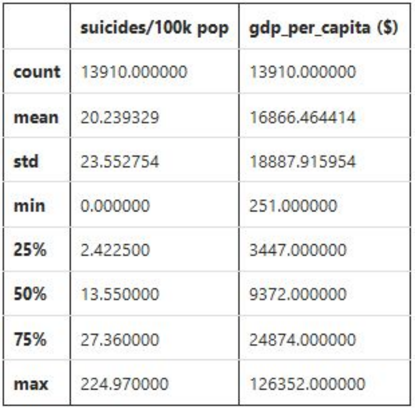


	suicides/100k pop	gdp_per_capita ($)
count	27820.000000	27820.000000
mean	12.816097	16866.464414
std	18.961511	18887.576472
min	0.000000	251.000000
25%	0.920000	3447.000000
50%	5.990000	9372.000000
75%	16.620000	24874.000000
max	224.970000	126352.000000

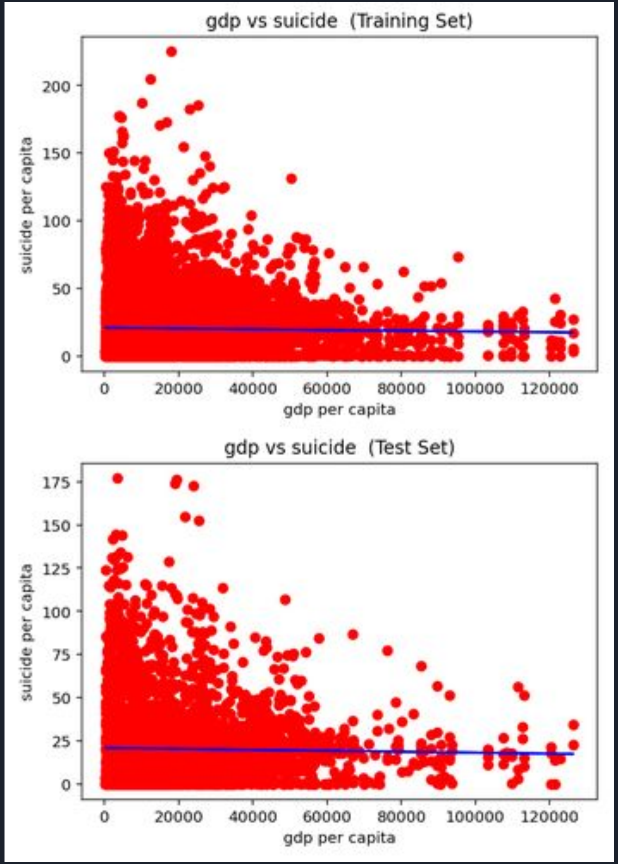


Male Linear Regression

- Male suicide rates data set information
- Model's Coefficient 
```
[-2.76333083e-05]
```
- Model's Intercept 
```
20.91781478554566
```
- R-Squared 
```
model.score(X, Y)
1.53970491556521e-05
```

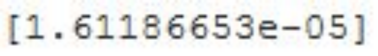


	suicides/100k pop	gdp_per_capita ($)
count	13910.000000	13910.000000
mean	20.239329	16866.464414
std	23.552754	18887.915954
min	0.000000	251.000000
25%	2.422500	3447.000000
50%	13.550000	9372.000000
75%	27.360000	24874.000000
max	224.970000	126352.000000



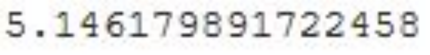
Female Linear Regression

- Female suicide rates data set information
- Model's Coefficient



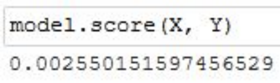
```
[1.61186653e-05]
```

- Model's Intercept

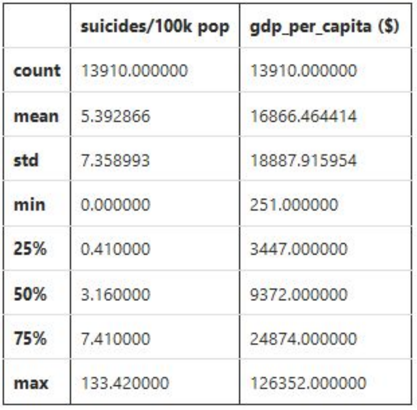


```
5.146179891722458
```

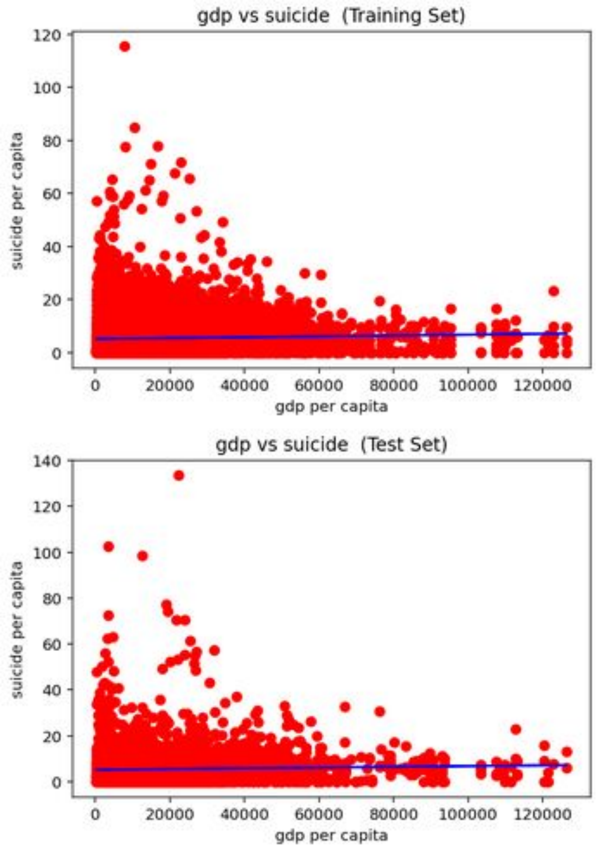
- R-Squared



```
model.score(X, Y)
0.002550151597456529
```



	suicides/100k pop	gdp_per_capita ($)
count	13910.000000	13910.000000
mean	5.392866	16866.464414
std	7.358993	18887.915954
min	0.000000	251.000000
25%	0.410000	3447.000000
50%	3.160000	9372.000000
75%	7.410000	24874.000000
max	133.420000	126352.000000



Linear Regression Algorithm Results

- Countries with less than $20000 GDP represents 70% of our dataset.
- No significant change in suicide rates as GDP increases
- Men have a greater suicide rate than women

Resources

https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

https://www.fastcompany.com/90349777/gender-inequity-costs-the-united-states-2-trillion-in-lost-gdp

https://ourworldindata.org/suicide

https://www.who.int/teams/mental-health-and-substance-use/suicide-data

https://databank.worldbank.org/source/world-development-indicators

Any Questions?